OPEN

Effects of Selective Exclusion of Patients on Preterm Birth Test Performance

J. Jay Boniface, PhD, Julja Burchard, MS, and George R. Saade, MD

The need to reduce the rate of preterm delivery and the recent emergence of technologies that measure hundreds of biological analytes (eg, genomics, transcriptomics, metabolomics, proteomics; collectively referred to as "omics approaches") have led to proliferation of potential diagnostic biomarkers. On review of the literature, a concern must be raised regarding experimental design and data analysis reporting. Specifically, inaccurate performance has often been reported after selective exclusion of patients around the definition boundary of preterm birth. For example, authors may report the performance of a preterm delivery predictor by using patients who delivered early preterm compared with deliveries at 37 weeks of gestation or greater. A key principle that must be maintained during the development of any predictive test is to communicate performance for all patients for whom the test will be applicable clinically (ie, the intended-use population), which for prediction of preterm birth includes patients delivering throughout the spectrum of gestational ages, as this is what is to be predicted, and not known at the time of testing. Using biomarker data collected from the U.S.-based Proteomic Assessment of Preterm Risk clinical trial, we provide examples where the area under the receiver operating characteristic curve for the same test artifactually improves from 0.68 (for preterm delivery at less than 37 weeks of gestation) or 0.76 (for preterm delivery at less than 32 weeks of gestation) to 0.91 when patients who deliver late preterm are excluded. We review this phenomenon in this commentary and offer recommendations for clinicians and investigators going forward.

FUNDING SOURCE: Sera Prognostics.

(Obstet Gynecol 2019;134:1333–8)

DOI: 10.1097/AOG.0000000000003511

Preterm delivery, which refers to delivery before 37 weeks of gestation, affects 15 million neonates born each year and varies from approximately 5–18% of all births across different geographies worldwide.[1] In the United States, it is the leading cause of neonatal death and the second leading cause of death in children before age 5 years. Preterm delivery is also a major source of long-term health consequences, including chronic lung disease, hearing and visual impairments, and neurodevelopmental disabilities such as cerebral palsy. The health-economic effects of preterm delivery in the United States was estimated to be between $26 and $31.5 billion[2,3] and costs continue to rise in most countries.[4]

Obstetric care providers routinely evaluate risk of preterm delivery using prior pregnancy history and cervical length, the two strongest traditional predictors of subsequent spontaneous singleton preterm delivery. Unfortunately, calculations based on published data[5,6] reveal that the risk factor of prior spontaneous preterm delivery is present in only 11% of all singleton pregnancies that result in spontaneous preterm delivery. Furthermore, calculations based on data from Hassan et al[7] indicate cervical length, as an independent predictor for spontaneous preterm delivery, only provides an additional attributable risk

From Sera Prognostics, Inc, Salt Lake City, Utah; and the Department of Obstetrics & Gynecology, University of Texas Medical Branch, Galveston, Texas.

The analyses described in this manuscript were supported in full by Sera Prognostics, Inc.

The authors thank Drs. Gregory Critchfield, Durlin Hickok, Michael Gravett, Garrett Lam, Paul Kearney, Lawrence Weir, Nathan Price and Ashoka Polpitiya and Mr. Max Dufford for direction and editorial assistance.

Each author has confirmed compliance with the journal's requirements for authorship.

Corresponding author: George R. Saade, MD, Department of Obstetrics & Gynecology, University of Texas Medical Branch, Galveston, TX 77555-0587; email: gsaade@utmb.edu.

Financial Disclosure

J. Jay Boniface and Julja Burchard are employees of Sera Prognostics, a company that has developed a screening test for preterm birth prediction and continues to develop such tests. The other author did not report any potential conflicts of interest.

ISSN: 0029-7844/19

of 6%. Although racial disparities and risk factors, such as low socioeconomic status, maternal age, and low maternal body mass index (BMI), have been identified,[8,9] up to 50% of all preterm deliveries occur in women without any evident risk factors.[10] Clearly, there is a need for improved prediction of this serious health condition.

Interest in assessing the risk of preterm delivery and the development of technologies that measure hundreds of biological analytes (eg, genomics, transcriptomics, metabolomics, proteomics; collectively referred to as "omics approaches") have greatly increased the discovery of potential predictive biomarkers. However, biomarker predictive performance is sometimes determined after selective exclusion of cases adjacent to the clinical definition boundary for preterm delivery. For example, authors may report the performance of a preterm delivery predictor by using patients who delivered early preterm compared with deliveries at 37 weeks of gestation or greater. Best practices for development of omics tests were formatted into guidelines by the National Academy of Medicine's Committee on the Review of Omics-based tests in 2012.[11] Amongst many elements described in this authoritative publication was the requirement of demonstrating test performance in the intended-use population, which in the context of preterm delivery prediction covers patients destined for delivery at all gestational ages after screening.

CHARACTERISTICS OF PREDICTIVE TESTS

The receiver operating characteristic (ROC) curve formed from the sensitivity and specificity along the continuum of possible test scores provides a good representation of the predictive characteristics of a test (Fig. 1C, F, and I).[12] The area under the ROC curve (AUC) represents the overall predictive ability of the test, with an AUC of 0.5 indicating no predictive ability and an AUC of 1.0 representing perfect predictive ability. When developing such a ROC curve for pregnancy outcomes such as preterm delivery, investigators perform the test on a cohort of women and then follow them until delivery. Once all the enrolled patients deliver, the investigators divide the patients into those with the outcome (case participants) and those without the outcome (noncase participants). All of the enrolled patients who were not lost to follow-up should be categorized as either case or noncase participants. When developing predictive tests for preterm birth earlier than 37 0/7 weeks of gestation, some studies have reported biomarker and algorithm performance after selective exclusion of patients adjacent to the definition boundary of 37 weeks of gestation. For example, some investigators have published[13–18] or presented (Weiner et al. Futureborn^{TM}–prediction of future preterm birth <33w and preeclampsia/eclampsia <34w by 16w using a novel test in asymptomatic women. Am J Obstet Gynecol 2017;216:S196 [abstract]) test performance by comparing term delivery with preterm deliveries before an early gestational age cutoff (eg, less than 32, 34 or 35 weeks of gestation), or by omitting early term deliveries (eg, 37 and 38 weeks of gestation).[16] Another more subtle form of gapping can also be found in the recent report by Jelliffe-Pawlowski et al,[16] where very early preterm deliveries (less than 32 weeks of gestation) were included at an unnatural equivalent proportion relative to late preterm deliveries (32–36 weeks of gestation). The resulting study distribution of preterm births by gestational age week includes an unnatural flattening in birth rate creating in essence a partial gap. The issue that must be realized, however, is that such limitations of gestational age within the study population (eg, less than 32, 34, or 35 vs greater than 37 weeks of gestation) necessarily exclude a whole group of patients and their outcomes, which artificially inflates the apparent test performance metrics (eg, AUC). To illustrate the effect of gapping on test performance, we used actual biomarker data from a previous study[19] and simulated diagnostic performance with and without gapping.

ROLE OF THE FUNDING SOURCE

Sera Prognostic data and analytical support were used to bring attention to this important issue of gapping. Each author participated in conceptualizing the ideas in this manuscript, designing the analysis, drafting the manuscript, editing, and approving the final, submitted version. Ms. Burchard performed the statistical analyses. Each author declares that Good Publication Practice (GPP3) guidelines have been maintained. Specifically, the authors had access to relevant aggregated study data and other information (such as study protocol, analytic plan and report, validated data table, and clinical study report) required to understand and report research findings. The authors take responsibility for the presentation and publication of the research findings, have been fully involved at all stages of publication and presentation development, and are willing to take public responsibility for all aspects of the work. All individuals included as authors and contributors who made substantial intellectual contributions to the research, data analysis, and publication or presentation development are listed appropriately. The role of the sponsor in the design, execution, analysis, reporting, and funding is

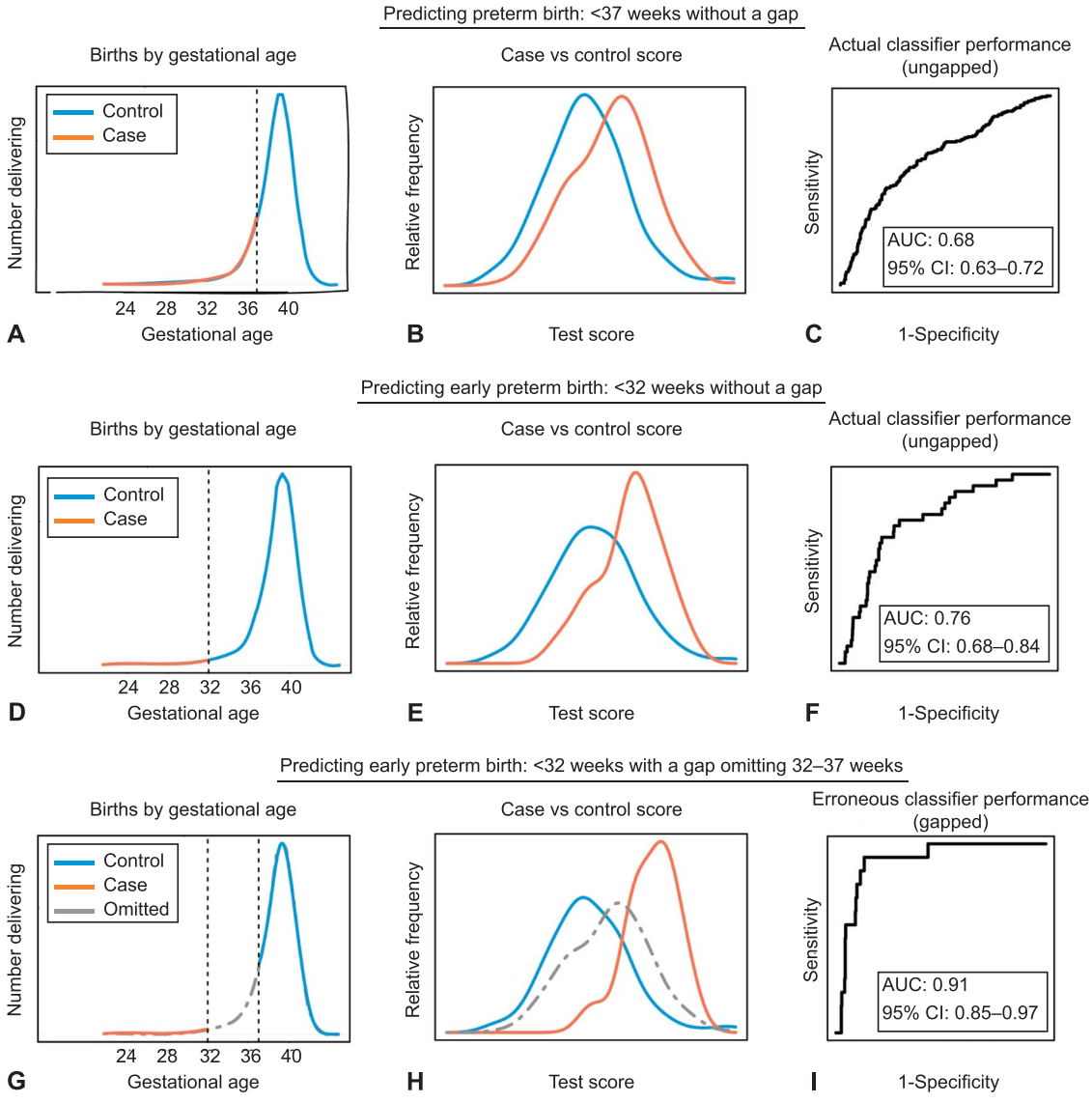

Fig. 1. Magnitude of erroneous estimation of test performance as a result of exclusion of patients. Shown are the distributions of gestational age at birth (**A**, **D**, and **G**), distributions of test scores by case–control status (**B**, **E**, and **H**), and corresponding actual, ungapped (**C**, **F**) or erroneous, gapped (**I**) test performance as estimated by area under the curve (AUC). **A**–**C**. All patients are included; case group, preterm birth at less than 37 weeks of gestation; control group, term birth at 37 weeks of gestation or greater. **D**–**F**. All patients are included; case group, preterm birth at less than 32 weeks of gestation; control group, births at 32 weeks of gestation or greater. **G**–**I**. Patients with gestational age at birth 32 weeks of gestation or greater through less than 37 weeks of gestation are excluded; case group, preterm birth at less than 32 weeks of gestation; control group, term births at 37 weeks of gestation or greater.

Boniface. Selective Exclusion in Preterm Birth Test Performance. Obstet Gynecol 2019.

fully disclosed. The authors' personal interests, financial or nonfinancial, relating to this research and its publication have been disclosed.

METHODS

Simulation of Test Performance

In a real-world unselected general population, the number of births increases with gestational age and peaks at full term (~40 weeks of gestation).[20] According to current practice and definitions, less than 37 0/7 weeks of gestation vs 37 0/7 weeks of gestation or greater is the dividing point by which preterm vs term births are defined (Fig. 1). Serum biomarker data, based on the ratio of insulin-like growth factor-binding protein 4 and sex-hormone binding globulin serum levels, were derived from analyses done on

blood drawn in weeks 19 and 20 of gestation from patients in the U.S.-based Proteomic Assessment of Preterm Risk clinical trial with singleton pregnancies not on progesterone after the 1st trimester and without signs or symptoms of labor at the time of blood draw. Simulations expanded a selected case–control study of 146 patients (41 preterm case participants and 105 control participants matched for distributions of BMI and gestational age at blood draw) by 5 times. The simulation process maintained the characteristics of the original data set with respect to gestational age and biomarker correlation and variability while increasing statistical power to 80% for detection of an AUC difference of 0.1 with $P<.05$ by DeLong's test,[21] a nonparametric approach to compare AUCs. The effects of an artificial gap on calculated performance metrics for a simulated test was modeled and is illustrated by the separation of the case participants' and control participants' test scores and by the corresponding AUCs. One thousand repetitions were performed, and a representative example was selected showing AUCs within their interquartile ranges for the proper intended-use population and for the same population where an artificial gap is created, with the difference between these AUCs at the median. Prevalence adjusted risk curves were generated from the intended-use population and from an artificially gapped population. A standard calibration plot was used to compare predicted and observed risk of preterm birth.[22] Analyses were performed in R 3.5.1 using the pROC package for AUC and the givitiR package for calibration plots.

Modeling Results

In the example data, the test demonstrates moderate performance using the current practice definition of less than 37 0/7 weeks of gestation vs 37 0/7 weeks of gestation or greater (Fig. 1B and C) and improved performance by lowering the boundary of case and noncase participants (less than 32 0/7 weeks of gestation vs 32 0/7 weeks of gestation or greater) and performing the analysis correctly without exclusion of patients (Fig. 1E and F). To illustrate the effect of gapping for a test intended to predict preterm delivery before 32 weeks of gestation, we then examined apparent performance with omission of births between 32 and 37 weeks of gestation (Fig. 1H and I). The omitted patients (Fig. 1G; hashed line) comprise approximately 8% of the total population and more importantly nearly 84% of all preterm births.[20] As illustrated in Figure 1H selective exclusion of patients widens the separation of case participants' and control participants' test scores. This results in an artifactual improvement in AUC to 0.91 (95% CI 0.85–0.97) for the gapped population compared with a correct AUC of 0.68 (95% CI 0.63–0.72) for preterm delivery at less than 37 weeks of gestation and a correct AUC of 0.76 (95% CI 0.68–0.84) for preterm delivery at less than 32 weeks of gestation in the proper intended-use population (Fig. 1C, F, and I). The difference in AUC in Fig. 1C compared with 1F does not show significance (DeLong's test, $P=.065$); differences in AUC for Fig. 1I compared with 1C or 1F are significant (DeLong's test, $P<.001$ and $P=.002$, respectively). The artifactual increase of 0.23 in AUC on gapping shown in Figure 1 is consistent with changes seen across 1,000 simulations (median 0.22, interquartile range 0.20–0.25). The effect on test performance can also be visualized using

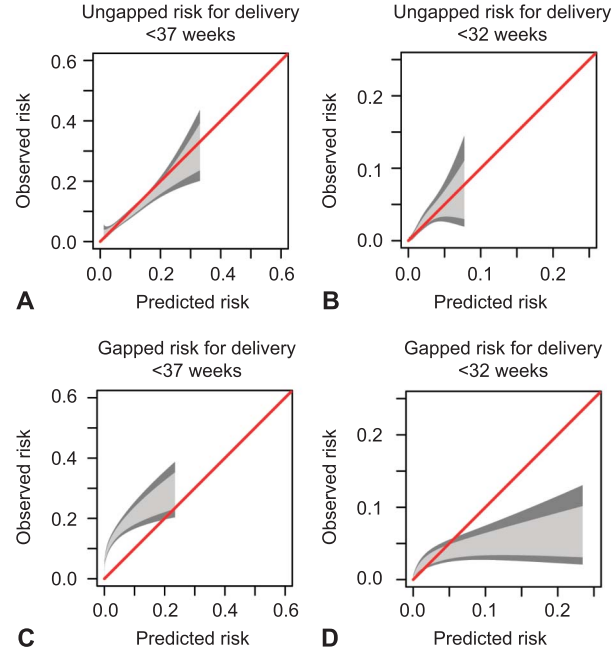

Fig. 2. Magnitude of agreement between predicted and observed risk as a result of exclusion of patients. Shown are predicted vs observed risks of preterm delivery when risks are calculated from an ungapped analysis (**A**, **B**) or gapped analysis (**C**, **D**), applied to a full intended-use population. **A.** Risk of preterm delivery at less than 37 weeks of gestation when all patients are included in test development. **B.** Risk of preterm delivery at less than 32 weeks of gestation when all patients are included in test development. **C.** Risk of preterm delivery at less than 37 weeks of gestation when patients with gestational age at birth 32 weeks of gestation or greater through less than 37 weeks of gestation are excluded in test development. **D.** Risk of preterm delivery at less than 32 weeks of gestation when patients with gestational age at birth 32 weeks of gestation or greater through less than 37 weeks of gestation are excluded in test development. *Red diagonal lines* represent perfect calibration of risk. The 80% and 95% CIs of the relationship between predicted and observed risk are represented by the width of the *light gray* and *dark grey shaded areas*, respectively.

Boniface. Selective Exclusion in Preterm Birth Test Performance. Obstet Gynecol 2019.

a calibration curve constructed from predicted and actual preterm delivery risk.[22] In such an analysis, the test is considered accurate when the predicted risk falls on the diagonal. On the other hand, when predicted risk falls above or below the diagonal the test underpredicts or overpredicts risk, respectively. Actual risk of preterm birth at less than 32 or less than 37 weeks of gestation is quite similar to the predicted risk when prediction is based on test scores of case and noncase patients representing the full intended-use population (Fig. 2A and B). However, predictions based on patients with an artificial gap in gestational age between case and noncase participants greatly underestimate the risk of preterm birth at less than 37 weeks of gestational age while overestimating by several fold the risk of preterm birth at less than 32 weeks of gestation at high test scores (Fig. 2C and D).

CLINICAL IMPLICATIONS

Recent progress in omics has provided exciting and novel opportunities for the development of innovative clinical tests. Amid this justified excitement, vigilance must be maintained by the scientific and clinical community in reporting and reviewing the conclusions of studies promoting biomarker performances. In this article, we address a concern that is critical to the validity of reports on biomarker performance for preterm delivery: the insertion of a gestational age gap in the study population does not allow for accurate estimates of predictive performance. The simulations based on actual data illustrate how "gapping" of the study population results in artifactual test performance for preterm birth prediction. Building estimates of AUC, sensitivity, specificity and predictive values in the context of selective exclusion of certain patients is inappropriate, because a prediction cannot be built for an intended-use population when it does not account for all such patients who will, in fact, exist in the population of patients to be tested. The performance of a test for the most severe preterm births should be determined by lowering the case and noncase boundary without omission of patients, as exemplified in Figures 1D, E, and F and 2B. The errors associated with "gapping" are not trivial and can have significant implications in both clinical practice and research. As exemplified here, gapped analyses may lead to overestimation of the test predictive abilities,

Table 1. Study Design and Analysis Considerations for the Test Characteristics to Be Clinically Applicable

Principle	Examples of Good Practice	Examples of Poor Practice
The outcome to be predicted (case status) and the comparison group (control status) must be appropriately defined, and study participants must be representative of the intended-use population (patients to be tested).	Case participants are defined as those who deliver before a specific gestational age (eg, 32, 35, 37 wk), and control participants (ie, noncase participants) are defined as those who deliver on or after the specific gestational age case definition without a gap.	Case participants are defined as those who deliver before a specific gestational age (eg, 32, 35, 37 wk), and control participants are defined as those who deliver on or after a later gestational age with a gap between case and control participants.
	The phenotypic distribution of case and control participants should match that of the intended-use population: eg, preterm delivery vs nonpreterm delivery (which includes any pregnancy complications other than preterm delivery, such as preeclampsia).	Case or control participants included in the analysis are enriched at unnatural distributions of gestational ages at birth (eg, very early preterm births and late preterm births are present at a 50:50 ratio, or control gestational ages peak earlier or later than occurs naturally). Case participants are defined as those with PTDs, and control participants are defined as those with term births without preeclampsia.
Data analysis is conducted in a manner to be reflective of application of the test to the defined intended-use population (patients to be tested).	Analyzing the tables of performance and the ROC curve, the total number of case and control (noncase) participants should be stated and should equal all the patients enrolled (with the exception of those lost to follow-up) who meet the intended-use population criteria, without exclusions based on times of delivery.	ROC curves and performance data are generated on a selected subset of patients who no longer represent the intended-use population.

PTD, preterm delivery; ROC, receiver operating characteristic.

which can lead to introduction of an ineffective test, overdiagnosis and unnecessary treatments, ultimately increasing cost and harm. Gapped analyses may be appropriate as proof of concept or for preliminary evidence to support further research, but such reports cannot imply clinical test performance nor be described as "clinical validation."

When evaluating preterm delivery prediction, it is important to clarify what is meant by a *control participant*. In such an analysis, control participant does not refer to a patient who has a normal pregnancy. A control participant is a patient who is not a *case participant*, that is, does not have the outcome being predicted. For the same reasons outlined above regarding gestational age gapping, it would be inappropriate to exclude patients who had a pregnancy complication (eg, preeclampsia) from the control group (or noncase group) when developing tests to predict preterm birth. To prevent any confusion, we suggest using case participant vs noncase participant to refer to those who have the outcome to be predicted and those who do not, rather than case participant and control participant.

When clinicians evaluate reported characteristics of any test to predict preterm delivery, we suggest following the checklist provided in Table 1. Although we focused on gestational age gapping in preterm delivery prediction, these principles apply equally to studies of other adverse outcomes in pregnancy that are influenced by gestational age, such as preeclampsia (early onset vs late onset), intrauterine growth disorders, and other complex maternal conditions that would mandate preterm delivery.

REFERENCES

1. Vogel JP, Lee AC, Souza JP. Maternal morbidity and preterm birth in 22 low- and middle-income countries: a secondary analysis of the WHO Global Survey dataset. BMC Pregnancy Childbirth 2014;14:56.
2. Behrman RE, Butler AS, editors. Preterm birth: causes, consequences, and prevention. Washington, DC: National Academies Press; 2007.
3. Caughey AB, Zupancic JA, Greenberg JM, Garfield SS, Thung SF, Iams JD. Clinical and cost impact analysis of a novel prognostic test for early detection of preterm birth. AJP Rep 2016;6:e407–16.
4. Reich ES. Pre-term births on the rise. Nature 2012;485:20.
5. McManemy J, Cooke E, Amon E, Leet T. Recurrence risk for preterm delivery. Am J Obstet Gynecol 2007;196:576–7.
6. Petrini JR, Callaghan WM, Klebanoff M, Green NS, Lackritz EM, Howse JL, et al. Estimated effect of 17 alpha-hydroxyprogesterone caproate on preterm birth in the United States. Obstet Gynecol 2005;105:267–72.
7. Hassan SS, Romero R, Vidyadhari D, Fusey S, Baxter JK, Khandelwal M, et al. Vaginal progesterone reduces the rate of preterm birth in women with a sonographic short cervix: a multicenter, randomized, double-blind, placebo-controlled trial. Ultrasound Obstet Gynecol 2011;38:18–31.
8. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet 2008;371:75–84.
9. Iams JD, Goldenberg RL, Meis PJ, Mercer BM, Moawad A, Das A, et al. The length of the cervix and the risk of spontaneous premature delivery. National Institute of Child Health and Human Development Maternal Fetal Medicine Unit Network. N Engl J Med 1996;334:567–72.
10. Iams JD. Clinical practice. Prevention of preterm parturition. N Engl J Med 2014;370:254–61.
11. Micheel C, Nass SJ, Omenn GS; Institute of Medicine (U.S.) Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials. Evolution of translational omics: lessons learned and the path forward. Washington, DC: National Academies Press; 2012.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
13. Cantonwine DE, Zhang Z, Rosenblatt K, Goudy KS, Doss RC, Ezrin AM, et al. Evaluation of proteomic biomarkers associated with circulating microparticles as an effective means to stratify the risk of spontaneous preterm birth. Am J Obstet Gynecol 2016;214:631.e1–11.
14. Esplin MS, Merrell K, Goldenberg R, Lai Y, Iams JD, Mercer B, et al. Proteomic identification of serum peptides predicting subsequent spontaneous preterm birth. Am J Obstet Gynecol 2011;204:391.e1–8.
15. Ezrin AM, Brohman B, Willmot J, Baxter S, Moore K, Luther M, et al. Circulating serum-derived microparticles provide novel proteomic biomarkers of spontaneous preterm birth. Am J Perinatol 2015;32:605–14.
16. Jelliffe-Pawlowski LL, Rand L, Bedell B, Baer RJ, Oltman SP, Norton ME, et al. Prediction of preterm birth with and without preeclampsia using mid-pregnancy immune and growth-related molecular factors and maternal characteristics. J Perinatol 2018; 38:963–72.
17. Ngo TTM, Moufarrej MN, Rasmussen MH, Camunas-Soler J, Pan W, Okamoto J, et al. Noninvasive blood tests for fetal development predict gestational age and preterm delivery. Science 2018;360:1133–6.
18. McElrath TF, Cantonwine DE, Jeyabalan A, Doss RC, Page G, Roberts JM, et al. Circulating microparticle proteins obtained in the late first trimester predict spontaneous preterm birth at less than 35 weeks' gestation: a panel validation with specific characterization by parity. Am J Obstet Gynecol 2019;220:488.e1–11.
19. Saade GR, Boggess KA, Sullivan SA, Markenson GR, Iams JD, Coonrod DV, et al. Development and validation of a spontaneous preterm delivery predictor in asymptomatic women. Am J Obstet Gynecol 2016;214:633.e1–24.
20. Martin JA, Hamilton BE, Osterman MJK, Driscoll AK, Drake P. Births: final data for 2016. Natl Vital Stat Rep 2018;67:1–55.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
22. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73.

PEER REVIEW HISTORY

Received May 18, 2019. Received in revised form July 12, 2019. Accepted July 25, 2019. Peer reviews and author correspondence are available at http://links.lww.com/AOG/B562.